37357 Advanced Statistical Modelling

Lab 2

Multiple Linear Regression and Transformations

Name: James Murray	**Student Number: 13879046**

The course library contains a simulated dataset called *salary* which contains the experience (in years), gender (dummy variable for male) and salary of 100 adults. Open SAS Studio and run the `libname` statement to load the data (see Lab Week 1 or Lab Week 2 if you have forgotten how to do this).

Produce a scatterplot to look at the relationship between salary and experience.

(1) Comment on the relationship between salary and experience.

There appears to be a positive correlation between salary and experience. As experience increases the variance in Salary appears to increase significantly. There is a 'splitting point' (heteroskedasticity) where two groups become apparent as experience passes 7 years – a higher wage bracket and a lower (almost stagnant) wage bracket.

Use the code below to produce a scatter plot that discriminates between genders.

```
proc sgplot data=mydata.salary;
      reg x=Experience y=Salary / group=Male;
run;
```

(2) Describe what the graph shows. Do you think gender should be included in the regression?

The splitting point highlighted above is the difference between gender in the Data. The increase in the Male group in wage is much faster than the increase in the Female group wages (steeper slope) on the regression line for the male group.

Noting the differences between the groups the ability to fit a model to the data will be highly dependent on the inclusion of gender in the model – it is more reflective of the training data set.

To test if there is an interaction between gender and experience, we first create an interaction variable to be included in the model. To create the new variable, run the following code.

```
data salary;
   set mydata.salary;
   Experience_Male = Experience*Male;
run;
```

Notes on the above code:

- The first line creates a new data set, `salary`. Since no library is specified, it will be created in your temporary WORK library.

- The second line copies all of the data from `mydata.salary` in your new dataset.
- The third line calculates the product of the variables `Experience` and `Male`, and saves it in a column called `Experience_Male`.

Now, run a multiple regression of salary on experience, gender and their interaction.
Hint: Remember from Lab Week 2 that you can generate additional diagnostic plots by using the `plot` *option in the* `proc reg` *statement.*

```
proc reg data=salary;
      model Salary=Experience Male Experience_Male;
run;
```

(3) Write down the regression equation and interpret the coefficients.

Salary = 40538 + Experience * 2070 + Male * -3060 + experience_male * 4187

For every year of experience wages go up $2070 and for males they go up an extra $4187 per year. This is a significant difference. The feature *Male* is not significant in the model's performance, but it will remain in the model so as not to fundamentally change the convergence of the male and female intercepts in the model.

(4) Comment on the fit of this model with a particular focus on any model violations identified by the residual plots. Comment on why a log transformation to the y-variable might help.

Violations – as the scale of the problem increases the spread of the variance increases indicating heteroscedasticity. this breaks one of the assumptions that underpins a lienar model. Variation increases as the both the wage values and years of experience increase.

A log transformation will reduce the variation at the higher end of experience across both male and female population groups thereby better suiting it to a linear model.

In order to address the issues with model fit, use the code below to take the log (base 10) transformation of the response variable.

```
data salary;
   set salary;
   Log_Salary = log10(salary);
run;
```

Run a regression on the transformed variable.

(5) Write down the regression equation and interpret the impact of the coefficients on the transformed and original salary scales.

Log_salary = 4.62335 + 0.01561*experience + 0.05121*male + 0.01499*experience_male

Interpretation:

Intercept is $42009.74

1 year of experience results in an increase in wages of 3.66%

Being a male results in a cumulative increase in wages of 12.5%

Having experience and being male results in a cumulative increase of 3.51%

(6) Comment on the fit of this model compared to the previous model and explain why a quadratic term may be required.

The R2 is particularly good for this model and all terms less male are significant in the model.

The log term has removed some of the heteroscedasticity from the model but it has highlighted the issue in the shape of the residuals for both experience and the experience_male variables. The inverted U shape in these two can be influenced using a quadratic variable - particularly at the higher end of the experience values.

Create a new variable for the quadratic term called `Experience_Sq`.
Hint: the quadratic term is essentially `Experience*Experience`

Refit the regression on the transformed response to include the quadratic term.

(7) Write down the regression equation and adjusted R^2. Interpret the coefficients on the original salary scale.

Adjusted R^2 = 0.9416

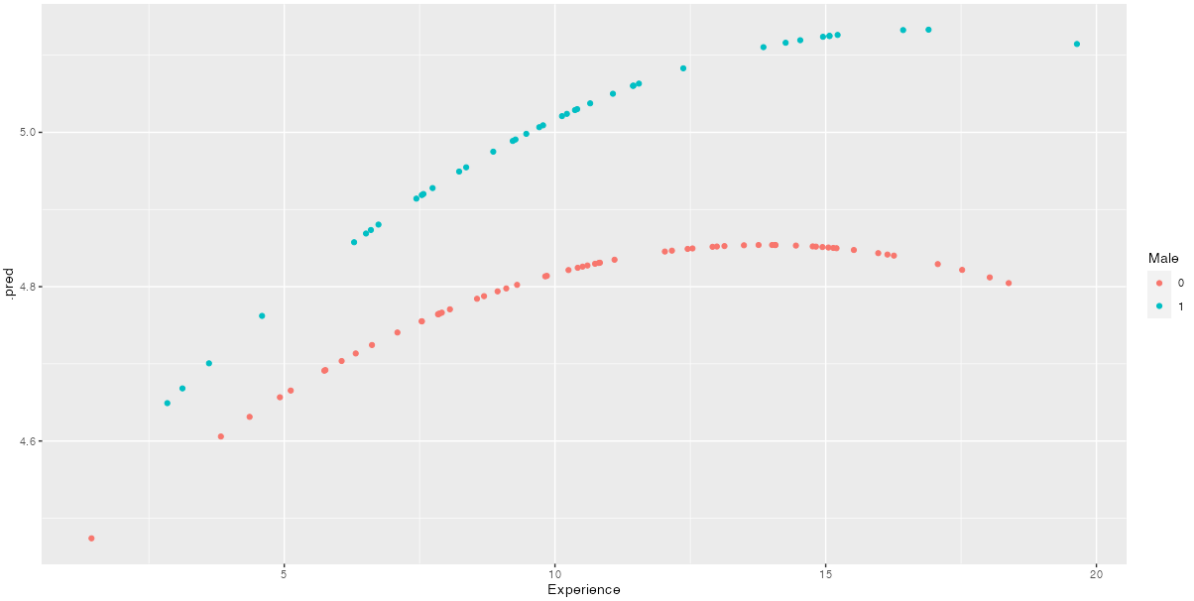
Intercept = $24036

1 year of experience results in a 17.0% increase

Being a male results in a further 12.9 % increase

Male experience increases a further 3.44% per year

Experience squared decreases the wage increase by -0.563%. This implies that at higher levels of experience the model levels off the experience with this negative. It will only impact the slope of the curve as the relative percentage increases at or around 17 years. The plot below shows this visually(using predicted values from the model) – Males are the Blue points and Females are the Red points. The effect occurs between 13 and 14 years of experience for females and between 16 and 17 years for males.


5.0
4.8
4.6
.pred
5
10
15
20
Experience
Male
0
1

(8) Comment on the fit of this model.

This model fits very well, possibly to well and may run the risk of over-fitting this set of Data. Another way to manage this model could be to remove the outliers. This may offer the model greater utility beyond this single data set when looking to make predictions.